Les modalités de l'évaluation

Denis Bonora

Édition électronique

URL : <https://journals.openedition.org/ries/3280>
DOI : 10.4000/ries.3280
ISSN : 2261-4265

Éditeur

France Education international

Édition imprimée

Date de publication : 1 septembre 1996
Pagination : 69-85
ISSN : 1254-4590

Référence électronique

Denis Bonora, « Les modalités de l'évaluation », *Revue internationale d'éducation de Sèvres* [En ligne], 11 | 1996, mis en ligne le 29 juillet 2013, consulté le 07 décembre 2022. URL : <http://journals.openedition.org/ries/3280> ; DOI : <https://doi.org/10.4000/ries.3280>

Ce document a été généré automatiquement le 29 septembre 2020.

Tous droits réservés

Les modalités de l'évaluation

Denis Bonora

- 1 Au cours de la période récente, les pratiques d'évaluation se sont diversifiées et ont occupé progressivement un éventail de plus en plus important de secteurs de la vie sociale et individuelle. De nouvelles formes d'évaluation se mettent en place, non seulement à l'égard des individus (par exemple dans le cadre de la formation continue, des centres de bilan et des procédures de reconversion), mais aussi des innovations pédagogiques (par exemple celles visant à la remédiation cognitive), ou encore des enseignants (sous l'angle de leur professionnalisme) et du système éducatif lui-même (dans ses multiples aspects, et notamment celui de son efficacité formatrice). Des procédures adaptées à chaque cas se différencient, tout en restant englobées, à côté des procédures traditionnelles, dans le concept générique d'évaluation, devenu de ce fait un concept « fourre-tout ». Dans le même temps, l'évaluation de l'élève, qui touche des générations entières d'individus pendant de nombreuses années, tend à perdurer sous des formes qui évoluent peu, dans la majorité des cas. Est-ce à dire qu'elle remplit bien son office ?
- 2 Nous nous cantonnerons ici à l'examen des procédures d'évaluation des apprentissages scolaires de l'élève, répondant à des objectifs du curriculum manifeste, intentionnel, à l'exclusion de ceux du curriculum latent. Ces procédures peuvent être invoquées pour répondre à des besoins variés, voire hétérogènes, qui nécessiteront des outils adaptés à chaque cas, et présentant des qualités métriques spécifiques.
- 3 Les types d'évaluation qui sont distingués classiquement se résument à trois grandes catégories : bilan sommatif, qu'il soit ou non à visée certificative ; évaluation pronostique, en vue notamment de l'intégration de l'élève dans un cycle d'étude nouveau ; et évaluation diagnostique, par exemple avec un objectif de régulation des apprentissages, cas dans lequel on la qualifie souvent de formative. Examinons d'abord les situations concrètes d'évaluation les plus typiques d'une perspective donnée. Nous réserverons pour la suite l'examen de la situation la plus couramment rencontrée, celle de l'évaluation dans le cadre de la classe, qui recouvre un ensemble hétérogène de pratiques. Nous chercherons alors à les caractériser, et dans ce cadre, les situations précédentes, plus « pures », pourront servir de points de repère.

L'évaluation sommative : l'apport de la docimologie

La situation d'examen

- 4 Elle s'applique en vue de l'octroi d'un diplôme ou d'une attestation de fin d'étude, de la collation d'un grade à l'issue d'un cycle de formation. Il s'agit là d'une situation typique de « bilan sommatif », basé sur la somme des connaissances que le sujet a pu accumuler au cours de cette formation. Ce type d'évaluation s'effectue à partir des performances du sujet.

La préparation des questions d'examen

- 5 Celle-ci consiste à échantillonner les éléments de connaissance contenus dans le cursus, de telle sorte que les questions posées, donc les performances observées, soient les plus représentatives possibles de ce cursus, tel qu'il apparaît, par exemple, dans les programmes officiels. Ce problème est difficile à résoudre, car on ne peut poser qu'un nombre limité de questions, ce qui empêche de couvrir tout le programme au moyen du questionnement ; il faudrait s'appuyer sur des corpus théoriques de nature psychologique et didactique pour définir les acquisitions les plus significatives du niveau attendu : une bonne réponse à telle question peut-elle être considérée comme révélatrice du niveau de structuration de la connaissance dans la discipline considérée ? C'est la difficile question de la hiérarchie chronologique des apprentissages - unique, ou bien multiple et vicariante - qui se trouve ici posée ; il n'est pas possible de déterminer avec certitude la valeur heuristique des acquis qu'on va contrôler, par rapport aux opportunités de réinvestissement ultérieur de ces acquis dans un contexte de vie personnelle et professionnelle. La première cause de cette difficulté est l'imprévisibilité des conditions de vie ultérieure de chaque candidat, et l'hétérogénéité de celles-ci, d'un individu à l'autre. Même si cette difficulté était surmontée, il resterait que ce choix des acquis à évaluer est orienté, chez le préparateur du questionnement, par sa propre conception de l'existence et son système de valeurs, paramètres qui varient de l'un à l'autre, ce qui se manifeste par des désaccords sur les objectifs de l'enseignement¹.
- 6 En outre, le questionnement ainsi construit devra présenter un niveau suffisant de congruence avec le type d'approche pédagogique et les systèmes conceptuels auxquels les candidats ont été exposés au cours de leur cursus. En particulier, le niveau de langage utilisé dans le questionnement ne devra pas faire écran.

La notation

- 7 Les diverses manifestations d'interactions candidat/examen sont maintenant bien connues : la performance du candidat pourra être en décalage par rapport à sa compétence, pour des raisons qui tiennent au caractère ponctuel de l'examen et à la spécificité de cette situation : l'état physique du sujet, sa motivation, son niveau d'éveil et de mobilisation pourront être différents de ce qu'ils sont habituellement. Les phénomènes de bachotage, d'« impasse » et de chance par rapport à la singularité des questions à traiter pourront également introduire des perturbations.
- 8 Par contre, il reste à examiner en particulier les conditions d'objectivité dans la notation. La qualité métrique qui est déterminante ici est la fidélité² : plusieurs juges qui

examineraient la même copie attribueraient-ils la même note à cette copie ? De nombreuses études docimologiques ont été réalisées et rapportées³ soit à partir de résultats d'examens réels, soit dans le cadre d'expérimentations, études dont les conclusions convergeaient en faisant ressortir que, d'un jury à l'autre, ayant corrigé les mêmes séries de copies ou des séries rendues équivalentes par tirage au hasard, les différences de moyennes étaient nettement supérieures à ce qu'on aurait pu attendre du fait des fluctuations aléatoires : le niveau d'exigence des différents jurys est donc inconstant et peut être en outre inconstant pour un même jury, d'une copie à l'autre ; les dispersions de notes étaient, elles aussi, différentes, en fonction, notamment, du degré de confiance de chaque juge dans sa propre procédure de notation (un juge peu confiant ayant en général tendance à réduire l'éventail des notes qu'il attribue), ce qui introduit une distorsion supplémentaire ; les corrélations entre séries de notes attribuées par des juges différents pour un même lot de copies étaient faibles, ce qui traduit des différences de classement des copies, d'un jury à l'autre : ces différences de classement sont la conséquence des désaccords entre juges en ce qui concerne les critères d'évaluation des copies, donc les objectifs de l'enseignement suivi antérieurement par les candidats. En outre, d'une matière à l'autre, la dispersion des notes varie de façon systématique, ce qui fausse le jeu des coefficients prévus pour les différentes matières de l'examen (une dispersion plus importante entraînant une sur-pondération).

- 9 La stabilité de la notation effectuée par un même professeur à deux moments différents sur une même série de copies n'est pas non plus parfaite et se détériore à mesure qu'on allonge le délai. Piéron⁴ rapporte, qu'après un délai de trois ans, l'instabilité observée est du même ordre que celle constatée entre deux juges évaluant au même moment.
- 10 Un autre problème concerne la cohérence entre épreuves écrites et orales pour une même discipline. Outre le fait que la corrélation entre notes est souvent éloignée de la valeur maximum, on a pu remarquer que le niveau d'exigence s'adoucissait de façon quasi-systématique quand on passe de l'écrit à l'oral, en raison notamment de la situation de communication directe dans ce dernier cas. Ce phénomène, manifeste au baccalauréat⁵, peut introduire un biais à l'avantage des élèves seulement admissibles (note moyenne entre huit et dix), à l'occasion de l'oral de rattrapage. Peut-être souhaitable au plan humain, il n'en reste pas moins regrettable au plan de l'objectivité, dans la mesure où ces épreuves orales ne sont appliquées qu'à une partie des candidats.
- 11 Toutes ces études docimologiques étaient basées sur le postulat d'une valeur « vraie » de la production de l'élève soumise à l'évaluation : la mesure de l'infidélité relative effectuée par les chercheurs à partir des distributions de notes visait alors à mettre en évidence l'erreur de mesure attachée à chaque procédure d'examen. Les solutions imaginées pour améliorer la notation étaient donc focalisées sur la préservation de la fidélité : concertation préalable entre correcteurs, double ou multi-correction, correction analytique à l'aide d'un barème pré-établi ou sur la base d'un corrigé-type, correction initiale collective de quelques copies, harmonisation des moyennes entre groupes, harmonisation des dispersions (par exemple, par le calcul de notes centrées-normées). Toutes ces procédures peuvent améliorer sensiblement l'objectivité de la note. Certaines sont utilisées pour le baccalauréat, telles que la correction analytique, éventuellement complétée par d'autres, mises en place par les commissions d'entente et d'harmonisation. Elles sont cependant encore loin d'être généralisées, car la plupart du temps, elles alourdissent sensiblement la tâche des évaluateurs.

Le contrôle continu

- 12 Celui-ci permet de pallier l'effet couperet de l'examen ponctuel (certes adouci, en ce qui concerne le baccalauréat, par la possibilité d'infléchir le verdict après consultation du livret scolaire). Il permet surtout de mieux échantillonner les apprentissages à évaluer, en particulier si l'on a pu établir un « référentiel de formation », qui répertorie les objectifs de la formation et les présente en termes de performances observables, dont la manifestation par l'élève témoignera que l'objectif a été atteint ou « maîtrisé ». Cependant, le contrôle continu présente d'autres inconvénients, dus en particulier au déficit de standardisation par rapport à la situation d'examen (qui elle-même, déjà, n'est pas un modèle à cet égard). En outre, il ne permet pas, surtout en l'absence de référentiel, d'échapper aux différences de niveau d'exigence des professeurs, qui connaissent probablement bien leurs élèves, mais moins bien ce que l'on peut attendre de l'ensemble de la population se trouvant au même point du cursus, ce qui les empêche d'ajuster leur niveau d'exigence.

La modération

- 13 La « modération » permet de ramener les évaluations à « une juste mesure » en situant la classe concernée, quant à sa moyenne et à sa dispersion, par rapport à la population d'ensemble des classes de même niveau dans le cursus. A cette fin, plusieurs procédures sont possibles⁶. On peut, par exemple, faire subir à la classe une épreuve standardisée dans la matière considérée, épreuve appliquée par ailleurs nationalement, dont les résultats, présentés dans le cadre d'un étalonnage national, serviront de points d'ancrage aux évaluations locales ultérieures. On peut aussi faire corriger quelques copies de chaque classe, à l'intérieur d'une circonscription donnée, par des évaluateurs extérieurs, les scores attribués servant alors de base de référence au professeur particulier pour situer les autres copies. Ces procédures de modération, qui sont utilisées couramment dans certains pays (Grande-Bretagne, Suède), restent assez méconnues en France. En outre, elles laissent entiers certains problèmes, tel que celui de l'effet d'« étiquetage » de certains élèves, sur lequel nous reviendrons.
- 14 Les différentes entorses à l'objectivité dans l'évaluation sommative, que nous venons d'évoquer, apparaissaient aux yeux des docimologues comme des erreurs aléatoires de mesure ou la manifestation de différences individuelles entre correcteurs.

Un modèle de fonctionnement de l'activité d'évaluation

- 15 Plus récemment, les travaux de Noizet et Caverni⁷, notamment, ont inauguré une nouvelle façon de concevoir le problème. Dans ce cadre, l'acte d'évaluation lui-même est un phénomène psychologique, soumis à ses lois propres qui affectent toute activité d'évaluation scolaire. Ces lois introduisent des variations systématiques dans la note attribuée, en fonction de la situation et non plus seulement des différences entre juges, variations qui sont préjudiciables à l'objectivité. Au cours des travaux expérimentaux qu'ils ont effectués en vue de construire un « modèle de fonctionnement » de l'activité d'évaluation, ils ont ainsi pu mettre au jour un certain nombre de sources de biais, qui commencent maintenant à être assez connues des enseignants.

Les effets séquentiels

- 16 Ces biais peuvent résulter d'effets séquentiels (dus à la succession des copies). Par exemple, si l'on fait noter une même série de copies par deux groupes de juges, en inversant l'ordre de ces copies d'un groupe à l'autre, on constate que les juges ont tendance à surestimer les copies de la première moitié par rapport à celles de la deuxième. Cependant, à l'intérieur de la série totale, d'autres tendances de biais se manifestent : les premières copies seront en général sous-évaluées, en raison de leur décalage par rapport au modèle préconçu de l'évaluateur. Par la suite, celui-ci aura tendance à intégrer les premières copies de la série dans son modèle de référence, et il deviendra moins exigeant, si bien que les suivantes pourront être relativement surévaluées. A un niveau plus local, une même copie ne sera pas notée de la même manière si elle vient après une bonne copie ou une mauvaise. Une copie de niveau donné sera surestimée si elle émerge d'un ensemble de copies de niveau médiocre, et inversement. Certaines copies pourront, par leurs caractéristiques, attirer l'attention du correcteur sur des critères de notation non pris en compte jusque-là.

Les effets « source »

- 17 Outre ces effets d'ordre, de contraste et d'ancrage, on relève également des effets « source » : la note attribuée par l'examineur est affectée par la connaissance qu'il a de la réussite antérieure de l'élève, de son statut scolaire, des caractéristiques de l'établissement auquel il appartient (effet de contamination). Certains traits caractérisant l'élève ou sa copie (qualité de l'écriture, fautes d'orthographe, etc.), pourront également influencer son jugement (effet de halo). Noizet et Caverni⁸ rapportent une expérience de notation en situation de simulation, qui montre bien la réalité de l'effet de contamination. Il s'agissait de la notation d'une série de versions anglaises par un groupe de professeurs. Ces copies étaient de niveau équivalent, mais portaient chacune mention d'une note fictive, censée avoir été obtenue antérieurement par le même élève. On a constaté que lorsque cette note antérieure était élevée, la note attribuée par les professeurs à la copie était en moyenne supérieure de deux points à celle attribuée aux copies portant mention d'une note antérieure faible. Ce phénomène de contamination, qui tend à « figer » le niveau reconnu de l'élève (effet d'étiquetage), peut être suffisant dans certains cas pour l'installer dans une situation d'échec. Les auteurs préconisent donc l'abandon des procédures courantes d'évaluation continue, par exemple en faisant effectuer les évaluations successives par des professeurs différents ou en proposant des épreuves à correction automatique (tests de connaissances).
- 18 Enfin, l'appartenance du sujet à une catégorie particulière peut, lorsqu'elle est connue de l'évaluateur, devenir la cause d'autres biais : par exemple, à niveau réel constant, la note attribuée varie en moyenne selon le sexe de l'élève, et surtout son âge et la PCS (profession et catégorie sociale) de ses parents⁹. Ces biais découlent d'un « effet d'attente » du professeur, qui tend alors à surestimer la production d'un élève appartenant à une catégorie à l'égard de laquelle il a un préjugé favorable¹⁰. Ce type de biais s'avère particulièrement « pervers » en ce qui concerne la PCS, car dans ce cas, il se cumule avec deux autres phénomènes qui handicapent les élèves issus de milieux défavorisés : d'une part, l'écart objectif de réussite entre PCS dans les apprentissages de connaissances ; d'autre part, la plus grande expertise de certains groupes - en général,

appartenant à des PCS favorisées - pour satisfaire aux normes formelles de l'institution et, ainsi, se valoriser.

Les causes du manque d'objectivité de la notation

- 19 Les différentes causes du manque d'objectivité de la notation dans l'évaluation sommative se ramènent en fait à deux grandes catégories : celles qui tiennent à la difficulté de définir les apprentissages les plus significatifs, ou les plus importants pour l'élève et la société ; celles qui introduisent des fluctuations parasites et nuisent ainsi à la fidélité.
- 20 Dans le premier ordre d'idées, les points à élucider sont de nature essentiellement idéologique au départ et portent sur les buts de l'éducation, dont découlent en bonne partie les objectifs pédagogiques. La conception « matricielle » de la discipline¹¹ y joue un rôle essentiel : en effet, son critère d'intelligibilité¹² évolue dans le temps à l'intérieur d'un système éducatif donné, déterminant massivement les contenus à enseigner et les méthodes adéquates pour ce faire.
- 21 Quant à la deuxième catégorie de causes d'erreur, celles-ci peuvent être en bonne partie éliminées, en principe, par l'utilisation de tests de connaissances, qui tend à l'heure actuelle à se développer en France et ailleurs¹³ : hier encore apanage exclusif des psychologues de l'éducation et de l'orientation, ils sont utilisés maintenant par certains enseignants dans leur classe et par certains établissements de formation (écoles d'ingénieurs, etc.), à des fins de recrutement.
- 22 L'atout majeur de ces instruments d'évaluation est leur standardisation. Celle-ci caractérise aussi bien les questions posées (souvent sous forme de QCM) que les conditions de passation (consignes, durée, etc.), la correction des réponses (score partiel imposé pour chaque réponse, catégorisée d'avance) et l'attribution d'un score total (étalonnage sur une population de référence). Les questions sont préparées avec soin sur une base explicite (les programmes officiels, par exemple) et en principe font l'objet d'une pré-expérimentation qui permet un polissage de l'outil. La fidélité des évaluations, qui peut alors facilement faire l'objet d'estimations quantifiées, est nettement plus élevée qu'avec les procédures traditionnelles. Ce type d'instrument ne constitue cependant pas une panacée, bien que certaines critiques qui lui sont faites ne soient pas fondées réellement. En effet, s'il est tout à fait capable d'évaluer des performances complexes de l'élève, par contre il se prête peu à l'évaluation de ses capacités d'expression (bien qu'il soit possible d'élaborer des items « à réponse construite ») : il n'évalue donc en général que ce qui était « prévu d'avance », ce qui limite son champ d'application et laisse peu de place à la reconnaissance de la créativité. En outre, il nécessite l'emploi de précautions quant aux risques de divulgation, cela afin d'éviter le bachotage. Enfin, étant orienté vers la performance, le produit de la démarche de l'élève, il est limité quant aux possibilités d'évaluation de la démarche d'accès à ce produit.
- 23 Il s'agit là d'instruments qui, s'ils sont construits dans les règles, apportent des garanties quant aux qualités psychométriques de l'évaluation et que les acteurs du système éducatif gagneraient à utiliser plus largement, sans toutefois en faire leur outil exclusif d'évaluation, option extrême qui les priverait de l'usage des autres fonctions possibles de l'évaluation traditionnelle, notamment comme stimulateur de la motivation extrinsèque¹⁴. Il convient aussi de se méfier du possible « effet en retour » des pratiques d'évaluation qui, d'une position aval par rapport au processus d'enseignement, peuvent tendre à glisser vers une position amont, et donc à remodeler les contenus enseignés, par un

ajustement plus ou moins conscient de l'enseignement dispensé sur la nature du « produit » final évalué.

L'évaluation pronostique : l'orientation, la sélection, le concours

Les outils du pronostic

- 24 De ces trois concepts, le premier - l'orientation - peut apparaître comme très hétérogène par rapport aux autres, dans la mesure où, telle qu'elle est mise en œuvre en France dans le dispositif institutionnel - notamment les centres d'information et d'orientation (CIO) -, elle se situe aux antipodes d'une procédure de sélection : on cherche au contraire, si possible dans le cadre d'une formation spécifique étalée dans le temps, à développer et ajuster son système de représentations, notamment en ce qui concerne ses propres intérêts et les caractéristiques attachées aux différents types d'emplois (ou métiers, ou fonctions). Cependant, le sujet n'échappe pas à la nécessité, à un moment ou l'autre, de mettre en parallèle les exigences de formation et, par suite, les capacités et autres caractéristiques individuelles nécessaires pour l'accès à ces emplois, et ses propres caractéristiques individuelles. On se trouve alors confronté - même si, dans ce cas, le sujet peut avoir un rôle moteur dans l'initiative de l'acte d'évaluation - à une problématique isomorphe à celle observée dans la mise en œuvre d'une procédure de sélection ou un concours d'admission : la question à élucider comporte toujours un problème de pronostic. Dans le premier cas, le sujet a besoin d'avoir une information sur sa probabilité d'adaptation et de satisfaction ultérieures dans la formation ou l'activité professionnelle envisagée, compte tenu de ses caractéristiques propres ; dans les autres, l'instrument utilisé devra permettre d'ordonner les candidats selon la dimension de probabilité de réussite ultérieure, en vue d'une sélection des plus « aptes ». On peut remarquer qu'alors la procédure est le plus souvent assortie d'un *numerus clausus* : il existe donc un seuil prédéterminé de réussite relative qui permet de trancher entre l'admission et le rejet, tandis que, dans le cas de l'orientation, ce seuil objectif n'existe pas.
- 25 En outre, la nécessité de sélectionner parmi les plus aptes amène à construire des outils qui discriminent bien dans la zone des bons niveaux, donc basés sur des objectifs « de perfectionnement ». Par contraste, un bilan sommatif d'examen de passage dans la classe supérieure, qui doit permettre de savoir si le sujet pourra suivre cet enseignement, devrait porter sur les pré-requis nécessaires à cette poursuite, identifiables comme « socle de connaissances », discriminant bien à l'intérieur de la zone des niveaux intermédiaires ou plutôt faibles.

La validité empirique

- 26 Dans tous les cas l'instrument utilisé - que ce soit une épreuve de type conventionnel, une épreuve standardisée de connaissances ou encore un questionnaire - devra présenter une qualité métrique spécifique : une bonne validité empirique ou pronostique (qu'il convient de distinguer des autres types de validité : de contenu et hypothético-déductive). Celle-ci atteste du lien souhaitable entre ce qu'elle mesure - le niveau de connaissance du sujet, ses intérêts dominants ou ses aptitudes -, qui joue le rôle de prédicteur, et le critère, par exemple la réussite dans la formation ultérieure ou l'emploi envisagés. La validité¹⁵ sera

donc d'autant mieux assurée que les performances évaluées par l'instrument seront en relation plus étroite avec celles observables dans la formation ou l'emploi ultérieurs.

- 27 Dans les concours classiques, le problème de l'estimation de la validité empirique est difficile à résoudre, car les épreuves changent d'une session à l'autre. On peut, certes, préserver un certain isomorphisme de l'instrument d'une session à l'autre et, ainsi, avoir une idée de la validité moyenne de la procédure, à partir de données recueillies sur plusieurs sessions successives. Cependant, il n'est possible d'évaluer la réussite ultérieure dans la voie considérée que pour les sujets qui ont été admis, ce qui biaise, en la réduisant, la corrélation avec le prédicteur.
- 28 En relation avec le mode de calcul de la validité empirique d'un instrument d'évaluation, il est utile de garder en mémoire une limitation importante quant à son application à un individu particulier. En effet, même quand la corrélation prédicteur-critère est élevée, la liaison n'atteint jamais, dans la pratique, un aspect linéaire : elle reste toujours probabiliste. Si bien qu'un pronostic basé sur une corrélation forte pourra être en même temps très fiable au niveau du groupe, tout en restant incertain au niveau individuel, dans la mesure où il ne peut être exprimé en termes de tout ou rien. Même informé d'une probabilité de réussite faible, tout individu peut donc légitimement espérer réussir « quand même », et il peut même être des cas dans lesquels cette information l'y aidera, en le mettant en position de challenger.
- 29 Enfin, on peut noter que la validité empirique d'un instrument n'est pas une qualité intrinsèque de celui-ci. D'une part, elle n'a, à l'évidence, de signification que par rapport au critère sur lequel on l'a estimée ; d'autre part, elle peut varier selon les caractéristiques de la population de sujets auxquels on applique l'instrument. À cet égard, signalons que dans le recrutement en vue de postes qualifiés impliquant une phase préliminaire de formation assez longue (cas dans lesquels les erreurs de pronostic sont donc coûteuses), on tend à prendre en compte le fait que des individus différents peuvent réussir de façon équivalente dans la variable-critère en utilisant des approches différenciées, en relation avec leurs caractéristiques personnelles : dans ce cas, une même épreuve pourra avoir une valeur pronostique élevée pour un sous-groupe donné et faible pour un autre. Une démarche d'élucidation pourra alors consister à rechercher les variables « modératrices »¹⁶ caractérisant ces groupes, qui permettront d'utiliser un modèle de pronostics multiples ou différentiels, basé sur une série d'épreuves pour lesquelles les pondérations varieront d'un sous-groupe à l'autre¹⁷. Ainsi, dans certains secteurs la fonction pronostique de l'évaluation évolue-t-elle en s'affinant.
- 30 Certains utilisent les notes scolaires pour formuler un pronostic de réussite à l'examen final. Or, compte tenu de la faible fidélité de ce type d'évaluation, on ne peut s'attendre qu'à une validité plutôt médiocre de tels pronostics : c'est effectivement le cas, puisque les coefficients de corrélation atteignent rarement la valeur 0,50 (ce qui équivaut à une amélioration sur le hasard de l'ordre de 15 %, dans le cas d'un pronostic dichotomique, donc grossier). À cet égard, une étude récente effectuée par Chatel (1994) sur des résultats du baccalauréat B de 1992, et portant sur 4 300 élèves dans cinq académies, confirme les études antérieures, en faisant état de corrélations plutôt faibles entre les moyennes annuelles de ces lycéens et leurs résultats au baccalauréat : de 0,34 pour la philosophie à 0,64 pour les mathématiques et la première langue étrangère.

L'évaluation diagnostique et l'évaluation formative

- 31 Dans un système où l'enseignement est dispensé collectivement, il apparaît naturel, *a priori*, que l'évaluation des apprentissages s'effectue sur la base des performances de l'élève, plutôt que sur ses démarches d'accès à celles-ci. En effet, les contraintes de la gestion d'un groupe rendent peu praticable l'observation individuelle que semble requérir l'évaluation de la démarche. En outre, l'évaluation est encore vue le plus souvent comme jouant exclusivement un rôle de contrôle *a posteriori*, donc extérieur au processus d'apprentissage lui-même.

La performance comme « norme » du groupe

- 32 Dans la classe, l'évaluation peut être utilisée pour établir de temps à autre, au cours de l'année scolaire, un bilan partiel reflétant la progression de l'élève, dans le cadre du programme prévu pour chaque discipline. Il s'agit là d'une évaluation qui s'apparente à l'évaluation sommative. C'est le cas, par exemple, des « épreuves communes », dont les résultats pourront être utilisés pour la répartition des élèves en groupes de soutien ou en « modules ». L'élève y est jugé individuellement à partir de sa performance. Le professeur pourra aussi utiliser les résultats groupés de sa classe pour s'auto-évaluer et réguler globalement son enseignement, en particulier le rythme de sa progression.
- 33 L'élève et ses parents pourront, quant à eux, disposer d'une information sur le niveau d'adéquation du travail scolaire consenti par l'élève, pour autant que le professeur aura fourni en même temps que la note individuelle des clés pour la décrypter : signification de notes-repères à l'intérieur de l'échelle utilisée ou, à tout le moins, note moyenne de la classe, des classes de même type dans l'établissement, niveau d'exigence pour le passage dans la classe supérieure, etc. Dans ce contexte, la note est interprétée plutôt comme un indicateur partiel du niveau global de l'élève dans la discipline ou de son niveau par rapport à une unité d'enseignement suivi (cours ou série de cours, définie par son unité thématique). Le modèle psychométrique, apparenté à ce type d'approche par le niveau, est un modèle plutôt normatif, selon lequel la performance individuelle prend son sens quand elle est resituée par rapport à son groupe de référence (c'est la fonction courante de l'étalonnage dans le cadre de l'utilisation d'un test de connaissances, notamment). Pour être pertinent, ce type d'évaluation doit avoir les mêmes qualités d'objectivité que les autres formes d'évaluation sommative évoquées précédemment et principalement la fidélité, outre la sensibilité, ou puissance de discrimination entre les bons et les moins bons, qui nécessite d'ajuster le niveau de difficulté du questionnement en fonction des capacités des élèves. Les critiques adressées précédemment à l'évaluation sommative s'appliquent ici pleinement et leur prise de conscience progressive par les enseignants ne peut que les inciter à une attitude modérée dans les extrapolations envisageables sur la base des notes observées (notamment celles débouchant sur des prises de décision quant à l'orientation des élèves).

La performance comme « critère » à satisfaire

- 34 Cependant, l'évaluation dans la classe peut être aussi conçue en mettant l'accent sur l'aide à l'enseignement en tant que démarche pédagogique. Partant d'une phase initiale

de définition de ses objectifs pédagogiques, formulés en termes de comportements observables de l'élève, l'enseignant mettra au point des exercices de contrôle adaptés à chaque objectif. La réussite de l'élève sur chacun de ces groupes d'exercices témoignera de son niveau de maîtrise de ces différents objectifs à mesure que se déroule le programme d'enseignement, ce qui suggérera les infléchissements utiles de la démarche ultérieure d'enseignement. Les objectifs peuvent être synthétisés dans un tableau appelé référentiel, qui les répertorie en même temps que les performances attendues de l'élève pour chacun d'eux. Cet usage du référentiel, qui a pris son essor dans l'enseignement technique français il y a une quinzaine d'années, où il tend à identifier des objectifs de savoir-faire et de compétences pluridisciplinaires plutôt que des savoirs disciplinaires, est encore peu répandu dans l'enseignement académique, où il est plus difficile à mettre en place en raison des résistances de ces disciplines à ce type d'analyse. Il est plus développé dans des matières où l'aspect moteur - facilement observable - des comportements est plus prégnant, comme l'éducation physique.

- 35 Ce type d'approche, d'inspiration néo-béhavioriste¹⁸, est encore basé sur la performance de l'élève, mais, cette fois, celle-ci est évaluée par comparaison à une performance attendue, qui est la définition opérationnelle de l'objectif visé. La performance individuelle prend son sens, ici, non plus par rapport au groupe de sujets (référence normative), mais par rapport à la performance-cible (référence critérielle). Le cadre de référence n'est donc plus interindividuel, mais devient intra-individuel : la performance de l'élève est évaluée en termes de progrès par rapport à sa propre performance antérieure et d'écart par rapport à la maîtrise de l'objectif comportemental.
- 36 Avec le glissement vers une telle conception de l'évaluation, la fidélité, qualité psychométrique jugée essentielle dans la perspective précédente comme indice d'approximation de la note « vraie », perd de sa prégnance au profit d'un souci accru d'adéquation par rapport au contenu de la discipline : la qualité privilégiée devient la validité de contenu. Celle-ci ne peut s'estimer quantitativement que de façon indirecte, par exemple par le degré d'accord entre spécialistes de la discipline sur la représentativité des objectifs évalués par l'instrument par rapport à ceux plus ou moins explicites de la discipline au niveau de cursus considéré. A ce sujet, on peut noter qu'en France, depuis une vingtaine d'années, les textes officiels ont donné une place importante à l'analyse des objectifs de l'enseignement des disciplines scolaires.
- 37 Ce type d'évaluation est aussi appelé évaluation formative dans la mesure où les résultats de sa mise en œuvre vont être réinvestis dans le processus d'enseignement, l'absence de maîtrise détectée chez certains élèves pour tel objectif pédagogique pouvant donner lieu à un retour en arrière, au cours duquel ces élèves réétudieront les unités de programme correspondantes, éventuellement selon une démarche d'approche différente.

L'évaluation de la démarche

- 38 Dans ce cadre, ce qui intéresse le formateur, ce n'est plus tant la performance du sujet en elle-même que la démarche qu'il utilise pour y parvenir. L'évaluation, de statique qu'elle était, devient dynamique dans la mesure où elle s'implique dans le déroulement de la démarche d'accès à la performance. Cette dernière n'apparaît plus alors que comme une manifestation partielle de la compétence cognitive du sujet, partielle parce que masquée par la spécificité de la nature du questionnement, le décalage éventuel des arrière-plans culturels entre l'élève et le questionnement, et l'existence fréquente de chemins

intellectuels alternatifs (vicariances) pour atteindre la « bonne » réponse. Les chemins sont de niveau et de portée heuristique différents, et le rôle du formateur est alors conçu comme un rôle d'aide au choix et à la mise en œuvre des chemins les plus heuristiques. A partir de là, on voit que l'évaluation n'est plus un moment séparé, un temps mort, dans le processus d'enseignement, mais qu'au contraire il s'y intègre. Dans l'évaluation formative axée sur la performance, évoquée précédemment, la régulation du processus pédagogique est de type rétroactif (par remodelage après coup). Or, elle est devenue ici interactive, car elle se manifeste « en temps réel » à travers un processus de négociation et une recherche d'explicitation menée conjointement avec l'élève. Un troisième niveau de régulation pourrait être défini, qui achève le passage d'une logique d'enseignement à une logique d'apprentissage, dans laquelle l'élève prend la plus grande partie de l'initiative : il met alors en œuvre, en effet, une démarche autorégulatrice, sur la base des informations que lui apporte le processus évaluatif. Cette approche, qualifiée d'évaluation formatrice, a été théorisée par Allai¹⁹ et expérimentée notamment par Nunziati²⁰.

- 39 Dans le contexte de l'évaluation formative, la réussite de l'élève dépend pour une part de sa compétence dans le domaine considéré, mais aussi de son fonctionnement métacognitif. L'évaluation formative dynamique intervient dans ce fonctionnement, avec un rôle d'aide au contrôle et au guidage de la tâche à accomplir. La réussite ainsi facilitée retentit sur les autres aspects métacognitifs de la démarche d'accès à la performance, notamment l'autoperception par l'élève de sa propre capacité, l'estime de soi et le degré de confiance dans sa propre démarche, ainsi que la prise de conscience de l'impératif de réussite, au-delà de l'impératif de simple réalisation de la tâche. Or, à l'heure actuelle, on prend de plus en plus conscience de l'importance capitale de l'estime de soi et de la confiance dans ses propres capacités, en tant que facteurs de la réussite, via le puissant moteur que constitue la motivation²¹. Ainsi se construit, nous semble-t-il, la boucle d'auto-entretien suivante : réussite > estime de soi > motivation intrinsèque²² > réussite²³.
- 40 Une telle conception de l'évaluation formative est moins ambitieuse que celle des psychologues de la mesure du potentiel d'apprentissage et de la remédiation cognitive, par exemple du type « programme d'enrichissement instrumental » ou « atelier de raisonnement logique »²⁴, qui vise à intervenir directement sur la structure de la compétence cognitive elle-même. Malgré cela, il va de soi que cette acception du terme « évaluation » est encore peu usitée dans les pratiques de terrain. Il y a, en effet, des difficultés importantes à surmonter, à la fois d'ordre conceptuel et matériel, pour réussir à mettre en œuvre des méthodes pédagogiques plurielles, dans un contexte de massification de la demande sociale d'éducation. Mais la tendance d'évolution est, elle, bien présente et soutenue par cet effort de théorisation. Notons cependant que l'enseignant désireux d'adopter une perspective d'évaluation formative dans sa pratique pédagogique ne pourra, ce faisant, se libérer complètement de la perspective sommative, ne serait-ce qu'en raison de l'utilité sociale de cette dernière – examens et concours – et de l'obligation qui lui est faite de « rendre compte », à l'administration de l'éducation nationale, aux parents et à la nation.

Le rôle de l'erreur dans l'évaluation de l'élève

- 41 Dans l'évaluation focalisée sur la performance, à visée sommative, l'erreur du sujet n'intervient en principe que dans la mesure où elle représente l'envers de la réussite, qui est seule « valorisée » : on s'intéresse au niveau de connaissances manifesté par l'élève et,

dans ce contexte, l'erreur n'est exploitée - négativement - que dans le cadre du décompte des points de score à attribuer à sa performance.

- 42 Dans l'évaluation focalisée sur l'élève et son apprentissage, l'erreur est au contraire considérée comme une source d'information permettant d'identifier la démarche de l'élève, et donc de lui fournir de façon personnalisée les éléments correctifs nécessaires pour assurer sa réussite dans l'apprentissage examiné. Cette valorisation de l'erreur, pratiquée à l'heure actuelle dans l'évaluation formative d'inspiration cognitiviste, n'est cependant pas une innovation : elle était déjà présente, par exemple, dans les années soixante, au cœur des évolutions de l'enseignement programmé : alors que la technique skinnerienne de base, de type linéaire, visait à une individualisation de l'apprentissage par l'adaptation au seul rythme de l'élève, l'introduction par Crowder du type de programme ramifié (ou à embranchements) amenait le sujet à répondre à la question posée sous forme de QCM, en choisissant parmi plusieurs possibilités. Chacune d'elle était prévue pour permettre l'identification d'un mode de fonctionnement différencié ayant entraîné l'erreur, et donc d'apporter au sujet les correctifs adaptés à son cas, au moyen d'un sous-programme spécifique qui pouvait à son tour être ramifié. Dès lors, l'individualisation ne portait plus seulement sur la vitesse d'apprentissage, mais aussi, de façon innovante, sur son contenu-même. Cette innovation a contribué par la suite au développement de deux perspectives : l'une, de nature psychologique, consistant à entrer dans la « boîte noire » des behavioristes, pour accéder, au delà du comportement observable du sujet, au processus cognitif qui y a conduit ; l'autre, d'inspiration pédagogique, visant les stratégies d'apprentissage des sujets et intégrant progressivement les facteurs personnels, conatifs²⁵ et sociaux. L'intégration de ces variables de conation dans les modèles de fonctionnement du sujet élaboré par les cognitivistes - intégration qui, il faut le reconnaître, pose quelque difficulté car ces variables ne se laissent pas facilement appréhender - devrait améliorer sensiblement leur portée heuristique, en les rapprochant de la réalité vécue « au quotidien », bref en les rendant plus « écologiques ».
- 43 A l'heure actuelle, la prise en compte de l'erreur n'est pas l'apanage exclusif des tenants de l'évaluation formative de type cognitiviste. On la voit souvent pratiquée, certes de façon non systématique et « périphérique », dans les procédures de correction appliquées par certains professeurs qui ne se réclament d'aucune obédience spécifique, par exemple au moyen d'annotations circonstanciées dans la marge des devoirs, qui renvoient l'élève à telle partie du cours, à tel principe méthodologique. Elle apparaît aussi, par exemple, dans les instruments développés par le ministère de l'Éducation nationale en vue des évaluations de masse : certains items font l'objet d'une cotation différenciée, qui permet d'identifier, au-delà de la bonne réponse, tel type d'erreur particulier. Il reste au professeur à se saisir de ces informations, puisqu'il dispose des résultats de ses propres élèves à l'issue de ces enquêtes, et de mettre en place des procédures localisées de remédiation, en rapport avec les types d'erreur manifestées par ceux-ci.
- 44 L'intérêt porté à l'évaluation des « connaissances » ne doit pas détourner complètement l'attention de l'enseignant de sa mission d'éducation qui comprend, certes, une mission d'instruction, mais en même temps la dépasse.
- 45 Il ne cesse donc pas, pour autant, de s'intéresser, au delà du sujet apprenant, à la personne de l'élève, dans tous les aspects du développement de sa personnalité et de son intégration à la vie en société. Cela implique un élargissement de la perspective didactique vers l'accès à une attitude « pédagogique » au plein sens du mot. A cet égard, on peut dire que la tendance d'évolution de l'évaluation - tout du moins quant à sa

théorisation, sinon encore tant dans ses pratiques -, de la performance à la démarche qui y conduit, et de l'apprentissage cognitif au métacognitif et au conatif, participe de cette préoccupation grandissante pour l'élève en tant qu'individu à part entière et, autant qu'il est possible, autonome.

NOTES

1. Denis Bonora, « Prémisse à l'évaluation : les préférences des professeurs à l'égard des objectifs pédagogiques », *L'Orientaion scolaire et professionnelle*, 17, 1988, n° 4, p. 323-342.
2. Dans la théorie classique de la mesure, la note observée d'un sujet est le résultat d'une composition additive de la note « vraie » évaluant sa capacité, et d'un terme d'erreur de mesure. Plus ce terme d'erreur est faible, et meilleure sera la fidélité, qui exprime ainsi le degré d'accord entre la note observée et la note « vraie ». L'un des indices qui permettent de l'estimer est la corrélation entre notes attribuées par des juges indépendants pour une même série de copies. La fidélité peut être aussi vue comme le degré d'accord entre résultats de deux passations successives d'une même épreuve (stabilité diachronique de la dimension mesurée) ou le degré d'accord entre parties d'une même épreuve (son homogénéité).
3. H. Piéron, *Examens et docimologie*, Paris, PUF, 1963.
- F. Bacher, « La docimologie », in : M. Reuchlin (éd.), *Traité de psychologie appliquée*, tome VI, Paris, PUF, 1973.
4. H. Piéron, op. cit.
5. M. Sery, « Le bac : une loterie. La docimologie et ses aléas », *Le Monde de l'Éducation*, n° 238, juin 1996, p. 38-39.
6. Bachet, op. cit. p. 68.
7. G. Noizet & J.-P. Caverni, *Psychologie de l'évaluation scolaire*, Paris, PUF, 1978.
8. G. Noizet & J.-P. Caverni, « Les procédures d'évaluation ont-elles leur part de responsabilité dans l'échec scolaire ? », *Revue française de pédagogie*, 1983, n° 62, p. 7-14.
9. M. Duru-Bellat & A. Mingat, « De l'orientation en fin de 5^e au fonctionnement du collège », in : *Cahiers de l'IREU*, 1988, n° 48.
10. C. Mathey-Pierre, « La différence institutionnalisée : pratiques d'orientation d'établissements scolaires et origine sociale des élèves », in : « De l'école à l'emploi. » *Cahiers du Centre d'étude de l'Emploi*, 1983, n° 26, Paris, PUF, p. 141-190.
11. M. Dévelay, « Didactique et pédagogie », in : « Éduquer et former », *Sciences Humaines*, 1996, hors série n° 12, p. 58-60.
12. Dévelay indique que la « matrice disciplinaire » à laquelle ce critère d'intelligibilité répond, par exemple pour la géographie, était hier - la géographie physique comme déterminant (de) la géographie humaine », alors qu'aujourd'hui, elle « emprunte à la nouvelle géographie, attentive aux stratégies, aux décisions et aux enjeux des acteurs pour aménager le territoire ». En ce qui concerne le français, la matrice relevait hier de la littérature, puis de la compréhension de la langue, pour aujourd'hui se centrer « sans doute sur la compréhension de la spécificité des formes d'écrits ».
13. Leur utilisation vient même d'être officialisée en Russie (janvier 96) en vue de l'admission dans certains établissements d'enseignement supérieur (*Le Monde de l'Éducation*, avril 1996, p. 56).

14. La motivation extrinsèque est celle qui n'est pas en rapport direct avec l'activité envisagée. Elle est favorisée par les renforcements, les notes, les gratifications matérielles ou symboliques. La motivation intrinsèque est au contraire celle qui oriente le sujet vers l'activité pour l'intérêt qu'elle présente en elle-même.

15. La validité empirique d'un instrument se calcule par un coefficient de corrélation* entre prédicteur et critère. Une corrélation éventuelle n'ayant des chances d'apparaître que si l'erreur de mesure n'est pas trop importante, on comprend que l'accès à une validité satisfaisante soit soumise à une condition préalable, celle d'un bon niveau dans la caractéristique métrique exigée déjà, on l'a vu, des examens : sa fidélité. La validité empirique peut s'opérationnaliser, lorsqu'on utilise un test, par une « table de prévision ». Celle-ci se présente comme un tableau à double entrée, où figure, pour chaque niveau atteint sur le prédicteur, la probabilité de réussite sur le critère, calculée sur un échantillon de sujets apparentés qui ont déjà subi cette épreuve (on peut ne pas se contenter d'une seule épreuve, mais en utiliser plusieurs, de nature différente, afin d'explorer les diverses facettes de prévision possibles).

* Cet indice statistique exprime la force de la liaison entre deux variables : depuis la valeur zéro, quand celle-ci est nulle, jusqu'à 1 quand elle est « fonctionnelle » (i.e. maximum), et affectée d'un signe (positif ou négatif) selon que la covariation est de même sens ou de sens inverse.

16. Il s'agit ici d'un concept différent de celui impliqué supra par les techniques de modération : une variable modératrice est une variable en interaction avec la variable prédictrice dans sa relation avec le critère. Par exemple, telle épreuve de concours n'aura pas la même validité pronostique selon qu'elle est subie par des sujets de milieu social ou d'ethnie A ou B.

17. J.-P. Fraïse, « La psychométrie à l'AFPA : son rôle dans l'orientation et le recrutement des demandeurs de formation professionnelle », *L'Orientation scolaire et professionnelle*, 1991, 20, n° 1, Spécial : L'évaluation individuelle dans ses divers cadres institutionnels, p. 127-141.

18. D. Bonora, « Les sciences de l'éducation et l'évaluation », in : J. Vogler (éd.), *L'évaluation : acteurs, pratiques, enjeux*, Paris, Hachette-Éducation, 1996 (sous presse).

19. L. Allai, « Régulations métacognitives : quelle place pour l'élève dans l'évaluation formative ? », in : Allai et al., *Évaluation formative et didactique du français*, Neuchâtel, Paris, Delachaux & Niestlé, 1993.

20. G. Nunziati, « Pour construire un dispositif d'évaluation formatrice », in : *Cahiers Pédagogiques*, 1990, n° 280, p. 47-64

21. R. J. Vallerand & E.E. Thill, *Introduction à la psychologie de la motivation*, Montréal, Études vivantes, 1993.

R. Viau, *La motivation en contexte scolaire*, Bruxelles, Montréal, De Boeck université, 1994.

R. Viau, « La motivation, condition essentielle de réussite », in : *Éduquer et former, Sciences Humaines*, 1996, hors série n° 12, p. 44-46.

22. Cf supra, la note n° 14.

23. Pour une discussion sur le sens de la causalité entre estime de soi et réussite, voir par exemple :

J.L. Muller, « La mise à l'épreuve d'un modèle des déterminants de la performance scolaire : l'utilisation de données méta-analytiques », *Bulletin de psychologie*, 1989, 42, n° 388, p. 114-120.

R.S. Newman, « Children's Achievement and Self-evaluation in Maths: a Longitudinal Study », in : *Journal of Educational Psychology*, 1984, 76, 5, p. 857-873.

24. M. Huteau et al., « Apprendre à apprendre : la question de l'éducabilité cognitive », in : G. Vergnaud (ed.), *Apprentissage et didactique, où en est-on ?*, Paris, Hachette-Éducation, 1994.

25. Le « conatif », qui réfère plutôt à la motivation du sujet, à ses intérêts, en liaison avec ses caractéristiques tempéramentales et leurs interactions avec le fonctionnement cognitif, peut être distingué de « l'affectif », qui réfère plutôt à ses émotions, ses sentiments en relation avec le vécu du sujet et son caractère agréable ou non.

RÉSUMÉS

Au cours des dernières décennies, les pratiques d'évaluation se sont diversifiées. L'auteur présente successivement les différents types d'évaluation des apprentissages scolaires. Chaque procédure a ses limites, répond à des objectifs variés, mais nécessite des outils adaptés qui doivent respecter des qualités d'application précises.

INDEX

Mots-clés : évaluation formative, évaluation prédictive, évaluation sommative, notation, évaluation de l'élève

AUTEUR

DENIS BONORA

Chercheur, Institut national d'étude du travail et d'orientation professionnelle, Paris, France